



# Differential meta-analysis of RNA-seq data from multiple studies

Andrea Rau, Guillemette Marot, Florence Jaffrezic

## ► To cite this version:

Andrea Rau, Guillemette Marot, Florence Jaffrezic. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics, 2014, 15 (1), pp.91. 10.1186/1471-2105-15-91 . hal-00978902

**HAL Id: hal-00978902**

**<https://inria.hal.science/hal-00978902>**

Submitted on 14 Apr 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

METHODOLOGY ARTICLE

Open Access

# Differential meta-analysis of RNA-seq data from multiple studies

Andrea Rau<sup>1,2\*</sup>, Guillemette Marot<sup>3,4</sup> and Florence Jaffrézic<sup>1,2</sup>

## Abstract

**Background:** High-throughput sequencing is now regularly used for studies of the transcriptome (RNA-seq), particularly for comparisons among experimental conditions. For the time being, a limited number of biological replicates are typically considered in such experiments, leading to low detection power for differential expression. As their cost continues to decrease, it is likely that additional follow-up studies will be conducted to re-address the same biological question.

**Results:** We demonstrate how  $p$ -value combination techniques previously used for microarray meta-analyses can be used for the differential analysis of RNA-seq data from multiple related studies. These techniques are compared to a negative binomial generalized linear model (GLM) including a fixed study effect on simulated data and real data on human melanoma cell lines. The GLM with fixed study effect performed well for low inter-study variation and small numbers of studies, but was outperformed by the meta-analysis methods for moderate to large inter-study variability and larger numbers of studies.

**Conclusions:** The  $p$ -value combination techniques illustrated here are a valuable tool to perform differential meta-analyses of RNA-seq data by appropriately accounting for biological and technical variability within studies as well as additional study-specific effects. An R package `metaRNASeq` is available on the CRAN (<http://cran.r-project.org/web/packages/metaRNASeq>).

**Keywords:** Meta-analysis, RNA-seq, Differential expression,  $p$ -value combination

## Background

Studies of gene expression have increasingly come to rely on the use of high-throughput sequencing (HTS) techniques to directly sequence libraries of reads (i.e., nucleotide sequences) arising from the transcriptome (RNA-seq), yielding counts of the number of reads arising from each gene. Due to the cost of HTS experiments, for the time being RNA-seq experiments are typically performed on very few biological replicates, and therefore analyses to detect differential expression between two experimental conditions tend to lack detection power. However, as costs continue to decrease, it is likely that additional follow-up experiments will be conducted to re-address some biological questions, suggesting a future

need for methods able to jointly analyze data from multiple studies. In particular, such methods must be able to appropriately account for the biological and technical variability among samples within a given study as well as for the additional variability due to study-specific effects. Such inter-study variability may arise due to technical differences among studies (e.g., sample preparation, library protocols, batch effects) as well as additional biological variability.

In recent years, several methods have been proposed to analyze microarray data arising from multiple independent but related studies; these meta-analysis techniques have the advantage of increasing the available sample size by integrating related datasets, subsequently increasing the power to detect differential expression. Such meta-analyses include, for example, methods to combine  $p$ -values [1], estimate and combine effect sizes [2], and rank genes within each study [3]; Hu *et al.* [4] and Hong and Breitling [5] provide a review and comparison of such

\*Correspondence: [andrea.rau@jouy.inra.fr](mailto:andrea.rau@jouy.inra.fr)

<sup>1</sup>INRA, UMR1313 Génétique animale et biologie intégrative, 78352 Jouy-en-Josas, France

<sup>2</sup>AgroParisTech, UMR1313 Génétique animale et biologie intégrative, 75231 Paris 05, France

Full list of author information is available at the end of the article

methods, and Tseng *et al.* [6] present a recent literature review and discussion of statistical considerations for microarray meta-analysis. In particular, Marot *et al.* [1] showed that the inverse normal  $p$ -value combination technique outperformed effect size combination methods or moderated  $t$ -tests [7] obtained from a linear model with a fixed study effect on several criteria, including sensitivity, area under the Receiver Operating Characteristic (ROC) curve, and gene ranking.

In many cases the meta-analysis techniques previously used for microarray data are not directly applicable for RNA-seq data. In particular, differential analyses of microarray data, whether for one or multiple studies, typically make use of a standard or moderated  $t$ -test [7,8], as such data are continuous and may be roughly approximated by a Gaussian distribution after log-transformation. On the other hand, the growing body of work concerning the differential analysis of RNA-seq data has primarily focused on the use of overdispersed Poisson [9] or negative binomial models [10,11] in order to account for their highly heterogeneous and discrete nature. Under these models, the calculation and interpretation of effect sizes is not straightforward. Kulinskaya *et al.* [12] recently proposed an effect size combination method for Poisson-distributed data, based on an Anscombe transformation, but this method is not well-adapted to RNA-seq data due to the presence of over-dispersion among biological replicates as well as zero-inflation. To our knowledge, no other transformation has been proposed to obtain effect sizes for over-dispersed Poisson or negative binomial data.

In this paper, we consider several methods for the integrated analysis of RNA-seq data arising from multiple related studies, including two  $p$ -value combination methods as well as a model fitted over the full data with a fixed study effect. We first demonstrate how the inverse normal and Fisher  $p$ -value combination methods can be adapted to the differential meta-analysis of RNA-seq data. Then we compare these two methods to the results of independent per-study analyses and a negative binomial generalized linear model (GLM) with a fixed study effect as implemented in the DESeq Bioconductor package [10]. All methods are compared on real data from two related studies on human melanoma cell lines, as well as in an extensive set of simulations varying the inter-study variability, number of studies, and biological replicates per study.

Finally, we note that our focus is on RNA-seq data arising from two or more studies in which all experimental conditions under consideration are included in every study (with potentially different numbers of biological replicates); differential analyses among conditions that are not studied in the same experiment are typically

limited, or even compromised, due to the confounding of condition and study effects.

## Methods

Let  $y_{gcrs}$  be the observed count for gene  $g$  ( $g = 1, \dots, G$ ), condition  $c$  ( $c = 1, 2$ ), biological replicate  $r$  ( $r = 1, \dots, R_{cs}$ ), and study  $s$  ( $s = 1, \dots, S$ ). Note that the number of biological replicates  $R_{cs}$  may vary between conditions and among studies. We use dot notation to indicate summations in various directions, e.g.,  $y_{gc \cdot s} = \sum_r y_{gcrs}$ ,  $y_{g \cdot \cdot s} = \sum_c \sum_r y_{gcrs}$ , and so on. Let  $\mu_{gcs}$  be the mean expression level for gene  $g$  in condition  $c$  and study  $s$ . For an integrated differential analysis of gene expression across all studies, two approaches can be envisaged: the combination of  $p$ -values from per-study differential analyses, and a global differential analysis. We illustrate both using the default methods and parameters of the DESeq (v1.10.1) analysis pipeline [10], although other popular methods, e.g., edgeR [11], could also be used; we note that the recent extensive comparison of Soneson and Delorenzi [13] provides a helpful guide to choosing an appropriate method and software package to use in practice.

## P-value combination from independent analyses

For the differential analysis of gene expression within a given study  $s$ , we assume that gene counts  $y_{gcrs}$  follow a negative binomial distribution parameterized by its mean  $\eta_{gcrs} = \ell_{crs} \mu_{gcs}$  and dispersion  $\phi_{gs}$ , where  $\ell_{crs}$  is a normalization factor to correct for differences in library size. A comparison of different methods to estimate  $\ell_{crs}$  may be found in Dillies *et al.* [14].

After obtaining per-gene mean and dispersion parameter estimates in each study independently, a parametric gamma regression is used to obtain fitted dispersion estimates by pooling information from genes with similar expression strengths. Subsequently, for each gene in each study, the null hypothesis to be tested is that there is no difference in the relative proportion of read counts attributed to each condition, or in other words, that the gene is non-differentially expressed. Per-gene and per-study  $p$ -values  $p_{gs}$  are computed using a conditioned test analogous to Fisher's exact test, where the  $p$ -value of a pair of observed count sums ( $y_{g1 \cdot s}, y_{g2 \cdot s}$ ) is calculated as the sum of all probabilities less than  $p(y_{g1 \cdot s}, y_{g2 \cdot s})$  given the overall sum  $y_{g \cdot \cdot s}$ :

$$p_{gs} = \frac{\sum_{\substack{a, b \geq 0 \\ a+b=y_{g \cdot \cdot s} \\ p(a, b) \leq p(y_{g1 \cdot s}, y_{g2 \cdot s})}} p(a, b)}{\sum_{\substack{a, b \geq 0 \\ a+b=y_{g \cdot \cdot s}}} p(a, b)}$$

where it is assumed that  $p(a, b) = p(a)p(b)$ , and  $p(a)$  and  $p(b)$  represent the probability of  $a$  and  $b$  counts in the first

and second conditions, respectively. These probabilities are calculated using the negative binomial distributions parameterized by the corresponding estimated mean and dispersion parameters,  $\mu_{gs}$  and  $\phi_{gs}$ .

Additional details are described by Anders and Huber [10] and in the DESeq package vignette. Once these vectors of raw  $p$ -values have been obtained for each study, we consider two possible approaches to combine them: the inverse normal and the Fisher combination methods. We note that both of these approaches assume that under the null hypothesis, each vector of  $p$ -values is assumed to be uniformly distributed.

#### Inverse normal method

For each gene  $g$ , we define

$$N_g = \sum_{s=1}^S w_s \Phi^{-1}(1 - p_{gs}) \quad (1)$$

where  $p_{gs}$  corresponds to the raw  $p$ -value obtained for gene  $g$  in a differential analysis for study  $s$ ,  $\Phi$  the cumulative distribution function of the standard normal distribution, and  $w_s$  a set of weights [15,16]. We propose here to define the study-specific weights  $w_s$ , as described by Marot and Mayer [17]:

$$w_s = \sqrt{\frac{\sum_c R_{cs}}{\sum_\ell \sum_c R_{c\ell}}},$$

where  $\sum_c R_{cs}$  is the total number of biological replicates in study  $s$ . This allows studies with large numbers of biological replicates to be attributed a larger weight than smaller studies. We note that other weights may also be defined by the user depending on the quality of the data in each study, if this information is available.

Under the null hypothesis, the test statistic  $N_g$  in Equation (1) follows a  $\mathcal{N}(0, 1)$  distribution. A unilateral test on the right-hand tail of the distribution may then be performed, and classical procedures for the correction of multiple testing such as the approach of Benjamini and Hochberg [18] may subsequently be applied to the obtained  $p$ -values to control the false discovery rate at a desired level  $\alpha$ .

#### Fisher combination method

For the Fisher combination method [19], the test statistic for each gene  $g$  may be defined as

$$F_g = -2 \sum_{s=1}^S \ln(p_{gs}), \quad (2)$$

where as before  $p_{gs}$  corresponds to the raw  $p$ -value obtained for gene  $g$  in a differential analysis for study  $s$ . Under the null hypothesis, the test statistic  $F_g$  in Equation (2) follows a  $\chi^2$  distribution with  $2S$  degrees of freedom. As with the inverse normal  $p$ -value combination

method, classical procedures for the correction of multiple testing [18] may be applied to the obtained  $p$ -values to control the false discovery rate at a desired level  $\alpha$ .

#### Additional considerations for $p$ -value combination

We note that the implementation of the previously described  $p$ -value combination techniques requires two additional considerations to be taken into account when dealing with RNA-seq data.

First, a crucial underlying assumption for the statistics defined in Equations (1) and (2) is that  $p$ -values for all genes arising from the per-study differential analyses are uniformly distributed under the null hypothesis. This assumption is, however, not always satisfied for RNA-seq data; in particular, a peak is often observed for  $p$ -values close to 1 due to the discretization of  $p$ -values for very low counts. To circumvent this first difficulty, as is commonly done for differential analyses in practice, we propose to filter the weakly expressed genes in each study, using the HTSFilter Bioconductor package [20] as described in the Additional file 1. We note that in so doing, it is possible for a gene to be filtered from one study and not from another. As will be seen in the following, this approach appears to effectively filter those genes contributing to a peak of large  $p$ -values, resulting in  $p$ -values that are roughly uniformly distributed under the null hypothesis.

Second, for the two  $p$ -value combination methods described above, unlike microarray data, under- and over-expressed genes are analyzed together for RNA-seq data. As such, some care must be taken to identify genes exhibiting conflicting expression patterns (i.e., under-expression when comparing one condition to another in one study, and over-expression for the same comparison in another study). In the case of microarray data, Marot *et al.* [1] suggested the use of one-tailed  $p$ -values for each study to avoid directional conflicts; as the inverse normal combination method was used in their work, the combined statistic thus follows a normal distribution, which is symmetric. Because under- and over-expressed genes may be found in the left and right tail, respectively, of the corresponding normal distribution, it is thus possible to use a two-tailed test to simultaneously study over and under-expressed genes. Note that Pearson [21] and Owen [22] proposed another alternative to handle conflicting differential expression if the Fisher combination method is used instead. However, in the case of RNA-seq data, the use of the conditioned test described above does not enable the separation of over- and under-expressed genes in distribution tails; this implies that it is not possible to use the approaches proposed by Marot *et al.* [1] or Owen [22]. We thus suggest that genes exhibiting differential expression conflicts among studies be identified post hoc, and removed from the list of differentially expressed genes; this step to remove genes with conflicting differential

expression from the final list of differentially expressed genes may be performed automatically within the associated R package metaRNASeq.

### Global differential analysis

For a global analysis of RNA-seq data arising from multiple studies, we assume that gene counts  $y_{gcrs}$  follow a negative binomial distribution parameterized by mean  $\eta_{gcrs} = \ell_{crs}\mu_{gcs}$  and dispersion  $\phi_g$ , where  $\ell_{crs}$  is the library size normalization factor. In order to estimate a possible effect due to study, a full and reduced model are fitted for each gene using negative binomial generalized linear models (GLM); the full model regresses gene expression on fixed effects for the experimental condition and study, while the reduced model regresses gene expression only on a fixed effect for the study.

Specifically, the full model is  $\log(\eta_{gcrs}) = \beta_g + \lambda_{gc} + \delta_{gs} + \log(\ell_{crs})$ , where  $\beta_g$  is an intercept,  $\lambda_{gc}$  is a fixed condition effect,  $\delta_{gs}$  a fixed study effect, and  $\lambda_{g1} = \delta_{g1} = 0$ , with the choice of the condition and study to be used as references being arbitrary. The reduced model is  $\log(\eta_{gcrs}) = \beta_g + \delta_{gs} + \log(\ell_{crs})$ . Per-gene dispersion parameters are estimated as before, where a parametric gamma regression is used to obtain fitted dispersion estimates by pooling information from genes with similar expression strengths across all studies.

We are now interested in testing the global per-gene null hypothesis

$$H_{0,g} : \forall c, \lambda_{gc} = 0 \text{ vs } H_{1,g} : \exists c \mid \lambda_{gc} \neq 0.$$

Following parameter estimation, the two models are compared using a  $\chi^2$  likelihood ratio test (with degrees of freedom equal to the number of conditions minus one) to determine whether including the experimental condition significantly improves the model fit. Note that for the global differential analysis we use the HTSFilter Bioconductor package [20] to filter the full set of data across studies prior to calculating  $p$ -values, resulting in a single vector of raw filtered per-gene  $p$ -values that may be corrected for multiple testing using classical procedures [18] to control the false discovery rate at a desired level  $\alpha$ . Additional details may be found in the DESeq package vignette.

## Results and discussion

### Application to real data

#### Presentation of the data

The negative binomial GLM and  $p$ -value combination methods were applied to a pair of real RNA-seq studies performed to compare two human melanoma cell lines [23]. Each study compares gene expression in a melanoma cell line expressing the Microphthalmia Transcription Factor (MiTF) to one in which small interfering RNAs (siRNAs) were used to repress MiTF, with three biological

replicates per cell line in the first (hereafter referred to as Study A) and two per cell line in the second (Study B). The raw read counts and phenotype tables for Study A are available in the Supplementary Materials of Dillies et al. [14], and the data from Study B from Strub et al. [23].

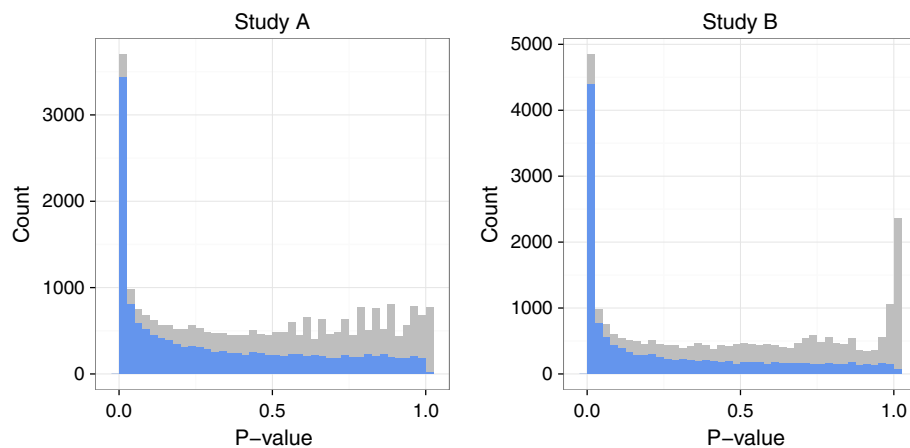
The characteristics of the data from these two studies are summarized in Additional file 1: Table S1. In particular, we note that the data from Study A tend to have larger total library sizes and a smaller number of unique reads (i.e. reads that appear once in the reference genome) than those from Study B; in addition, Study A appears to exhibit larger overall per-gene variability than does Study B (Additional file 1: Figure S8). These two points indicate that in this pair of studies, a considerable amount of inter-study heterogeneity appears to be present (Additional file 1: Figure S9).

### Results

After performing individual differential analysis for each study using the negative binomial model and exact test as described in the previous section, we obtained per-gene  $p$ -values for each study (Figure 1, histograms in background). As previously stated, an important underlying assumption of the  $p$ -value combination methods is that the  $p$ -values are uniformly distributed under the null hypothesis; we note that this is not the case here, especially for the second study, due to a large peak of values close to 1 resulting from the discretization of  $p$ -values. In order to remove the weakly expressed genes contributing to this peak in each study, we filtered the data from each study as proposed in Rau et al. [20], resulting in a distribution of raw  $p$ -values from each study that appears to satisfy the uniformity assumption under the null hypothesis (Figure 1, histograms in foreground).

The per-study filtered  $p$ -values were combined using the test statistics defined in Equations (1) and (2), and the corresponding results were compared to those of the intersection of independent per-study analyses and the global analysis using a negative binomial GLM with a fixed study effect as previously described. We note that for the independent per-study differential analyses, a gene is declared to be differentially expressed if identified in both studies with no differential expression conflict. For the Inverse normal and Fisher methods, a total of 310 (6.8% of differentially expressed genes) and 439 (9.0% of differentially expressed genes) genes were respectively found to exhibit conflicting expression between the two studies, and were subsequently removed from the final list of differentially expressed genes. Unsurprisingly, these genes tended to be those with relatively large  $p$ -values in both studies (Additional file 1: Figure S11).

In addition, we also investigated whether genes identified as differentially expressed by the Inverse normal and Fisher methods tended to be disproportionately



**Figure 1 Raw *p*-value histograms from per-study differential analyses of real data.** Histograms of raw *p*-values obtained from per-study differential analyses in the real data from Study A (left) and Study B (right): unfiltered (in grey) and filtered (in blue) using the method of [20]. Figure made using the ggp1ot2 package [24].

dominated by one study over the other, i.e. very small *p*-values in only one study (Additional file 1: Figure S12). Although Study B appears to have slightly more genes with very small *p*-values, for the most part, *p*-values for differentially expressed genes tend to be well-balanced between the two studies.

The Venn diagram presented in Figure 2 compares the lists of differentially expressed genes found for all methods considered. It may immediately be noticed that the independent per-study analysis approach is very conservative, and both of the *p*-value combination approaches (Fisher and inverse normal) considerably increase the detection power. In addition, a large number of genes are found in common among the *p*-value combination methods and the global analysis (3578 compared to only 1583 from the intersection of individual studies). In order to determine whether the genes uniquely identified by a particular method appear to be biologically pertinent, an Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com) was performed to identify functional annotation for the genes uniquely identified by the Fisher *p*-value combination method with respect to the global analysis, and vice versa. We note that the sets of genes uniquely identified by the Fisher method or the global analysis (Additional file 1: Tables S2 and S3), as well as the set of genes found in common (Additional file 1: Table S4), all appear to include genes of potential interest related to cancer or melanoma, which was the main focus of this set of studies. As such, for this pair of studies it appears that the union of genes identified by the two approaches may be of biological interest; to further study the effect of number of studies and inter-study variability on the performance of each method, we investigate an extensive set of simulated data in the following section.

### Simulation study

Data were simulated according to a negative binomial distribution,

$$Y_{gcs} \sim \mathcal{NB}(\mu_{gcs}, \phi_{gs})$$

where  $\mu_{gcs}$  and  $\phi_{gs}$  represent the mean and dispersion, respectively, for gene *g*, condition *c* and study *s*, and the mean-variance relationship is defined by

$$\text{Var}(Y_{gcs}) = \mu_{gcs} + \frac{\mu_{gcs}^2}{\phi_{gs}}.$$

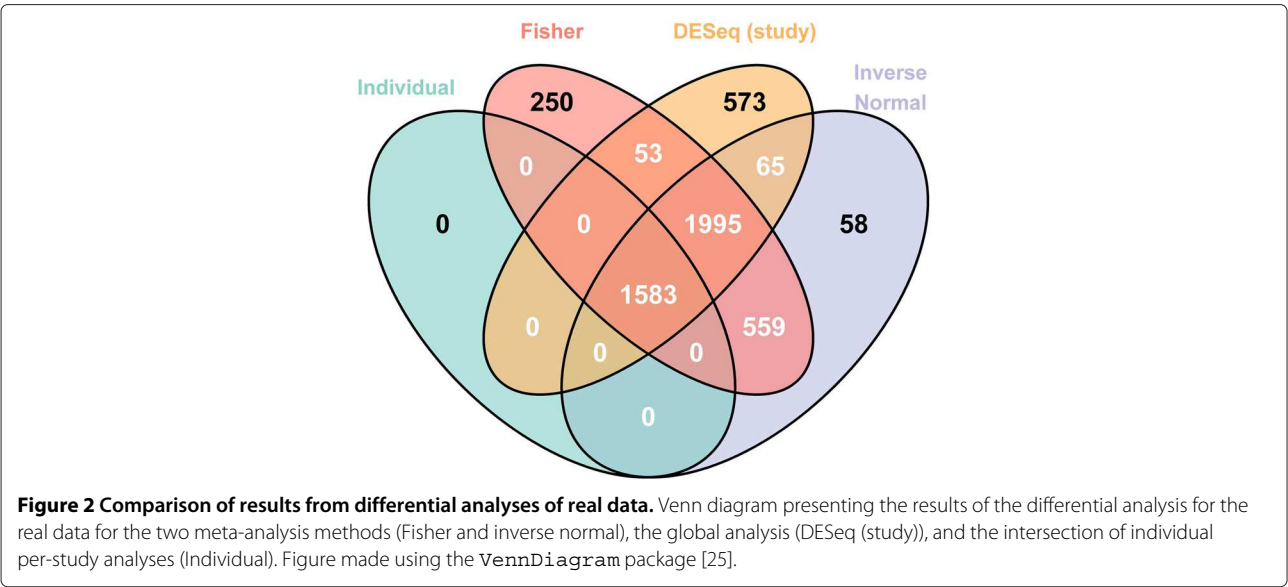
In order to incorporate inter-study variability, we consider the following situation for the mean parameter  $\mu_{gcs}$ :

$$\log(\mu_{gcs}) = \theta_{gc} + \varepsilon_{gcs}, \text{ and } \varepsilon_{gcs} \sim \mathcal{N}(0, \sigma^2),$$

where  $\theta_{gc}$  represents the mean for gene *g* in condition *c*,  $\varepsilon_{gcs}$  the variability around these means due to a study- and condition-specific random effect, and  $\sigma^2$  the size of the inter-study variability. Note that as  $\varepsilon_{gcs}$  affects  $\mu_{gcs}$  through a log link, the value of  $\exp(\varepsilon_{gcs})$  has a multiplicative effect on the mean.

### Parameters for simulations

To fix realistic values for the parameters  $\{\theta_{gc}, \phi_{gs}, \sigma\}$ , we first performed individual per-study differential analyses by fitting a negative binomial model with the default methods and parameters of the DESeq package on the unfiltered human data presented above. The per-study false discovery rate was subsequently controlled at the  $\alpha = 0.05$  level [18]. For the genes identified as differentially expressed in both studies,  $\theta_{g1}$  and  $\theta_{g2}$  were fixed to be the values of the empirical means (after normalization for library size differences) for each condition across studies. For the remaining genes, we set  $\theta_{g1} = \theta_{g2} = \theta_g$  to be the overall empirical mean (after normalization for library



size differences) for gene  $g$  across both conditions and studies. Using the gamma-family GLM fitted to the per-gene mean and dispersion parameter estimates for each study (Additional file 1: Figure S8), we fixed the dispersion parameter  $\phi_{gs}$  to be equal to the fitted values

$$\phi_{gs}^{-1} = \hat{\gamma}_{0s} + \frac{\hat{\gamma}_{1s}}{\theta_g},$$

where  $\hat{\gamma}_{0s}$  and  $\hat{\gamma}_{1s}$  are the estimated coefficients from the gamma-family GLM for study  $s$ , and  $\theta_g$  is the overall empirical mean for gene  $g$ . For weakly expressed genes, it has been observed that little overdispersion is present as biological variation is dominated by shot noise (i.e., the variation inherent to a counting process); for genes with  $\theta_g < 10$ , the dispersion parameter is therefore fixed to be  $\phi_{gs} = 10^{10}$ , which corresponds to nearly zero overdispersion (i.e., mean nearly equal to the variance).

Finally, the parameter  $\sigma$  is chosen to represent a range of values for the amount of inter-study variability. The observed human data exhibit a considerable amount of inter-study variability, corresponding to a value of roughly  $\sigma = 0.5$  (see Additional file 1: Figure S9). In the following simulations, four values are considered for the parameter  $\sigma$ :  $\{0, 0.15, 0.3, 0.5\}$ , representing zero, small, medium, and large inter-study variability, respectively. Finally, we note that for genes simulated to be non-differentially expressed, we set  $\varepsilon_{g1s} = \varepsilon_{g2s} = \varepsilon_{gs} \sim \mathcal{N}(0, \sigma^2)$ .

The simulation settings used for the number of studies and number of replicates per condition in each study are presented in Table 1 and were chosen to reflect the size of real RNA-seq experiments. When more than two studies were simulated, the same simulation parameters were used as for the first two, as determined from the real

data. For simplicity, the same number of replicates was simulated in each condition for all studies.

Methods and criteria for comparison

In addition to the intersection of independent per-study analyses (where genes were declared to be differentially expressed if identified in more than half of the studies with no differential conflict), the Fisher and inverse normal  $p$ -value combination techniques, and the global analysis with fixed study effect, we also considered a global analysis with no study effect. For each simulation setting and level of inter-study variability  $\sigma$ , 300 independent datasets were simulated, and the filtering method of Rau *et al.* [20] was applied, either independently to each study (for the independent per-study analyses and  $p$ -value combination techniques) or to the full set of data (for the global analysis).

For each method, performance was assessed using the sensitivity, false discovery rate (FDR) and area under the receiver operating characteristic (ROC) curve (AUC). In addition, we also considered a criterion to assess the “value added” for the  $p$ -value combination methods with

Table 1 Parameter settings for the simulations, including the number of studies and the number of replicates per condition in each study

Setting	# of studies	Replicates/study
1	2	(2,3)
2	3	(2,2,3)
3	5	(2,2,3,3,3)

Parameter settings for the simulations, including the number of studies and the number of replicates per condition in each study.



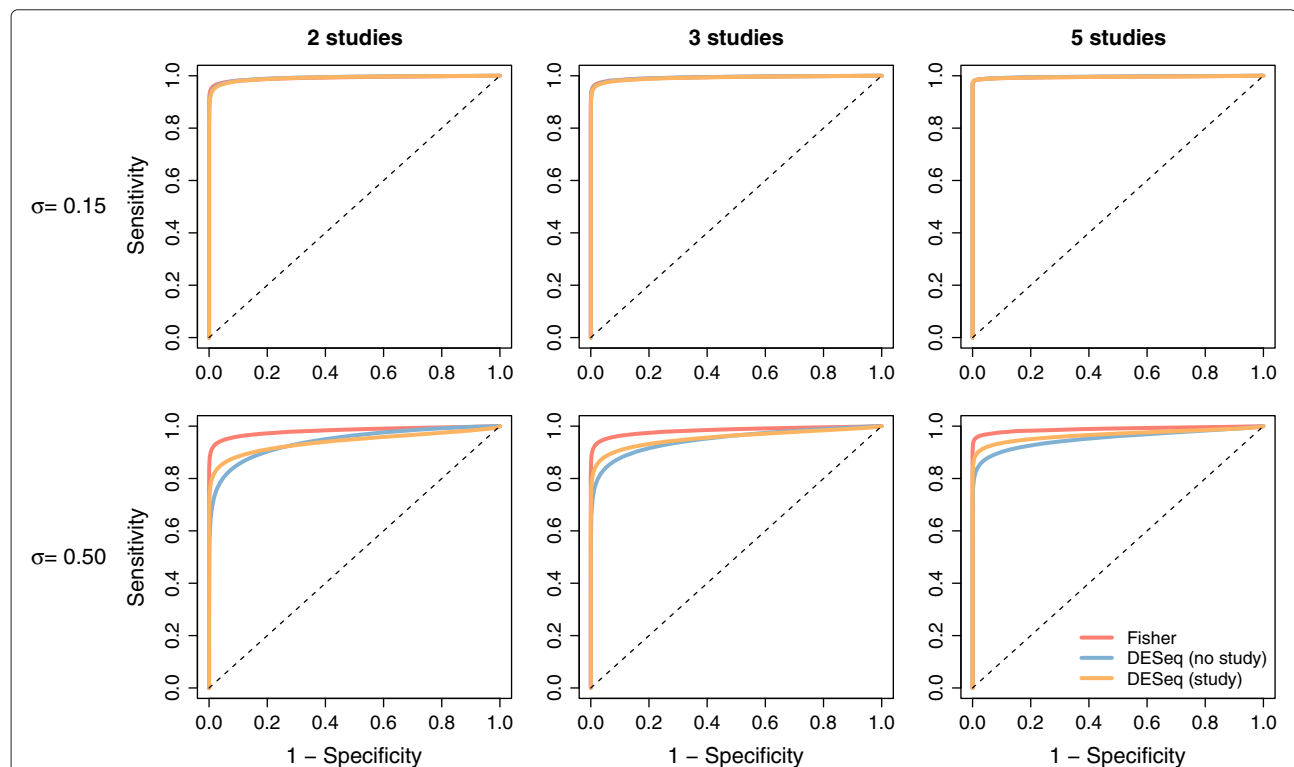
respect to the global analysis, and vice versa: the proportion of true positives among those uniquely identified by a given method (e.g., the Fisher approach) as compared to another (e.g., the global analysis).

## Results

The different methods were first compared with ROC curves, presented in Figure 3 for low and high inter-study variability (results for zero and moderate inter-study variability are shown in Additional file 1: Figure S5). We note that for clarity, the inverse normal method is not represented on these plots as its performance was found to be equivalent to the Fisher method. It can first be noted that for no or small inter-study variability ( $\sigma = 0$  or  $\sigma = 0.15$ ), no practical difference may be observed among the methods. On the other hand, for moderate to large inter-study variability ( $\sigma = 0.3$  or  $\sigma = 0.5$ ) differences among the methods become more apparent; this pattern is observed for any number of studies. As expected, including a study effect in the global analysis improves the performance over a naive global analysis without such an effect. We note that the two proposed meta-analysis methods (inverse normal and Fisher  $p$ -value combination)

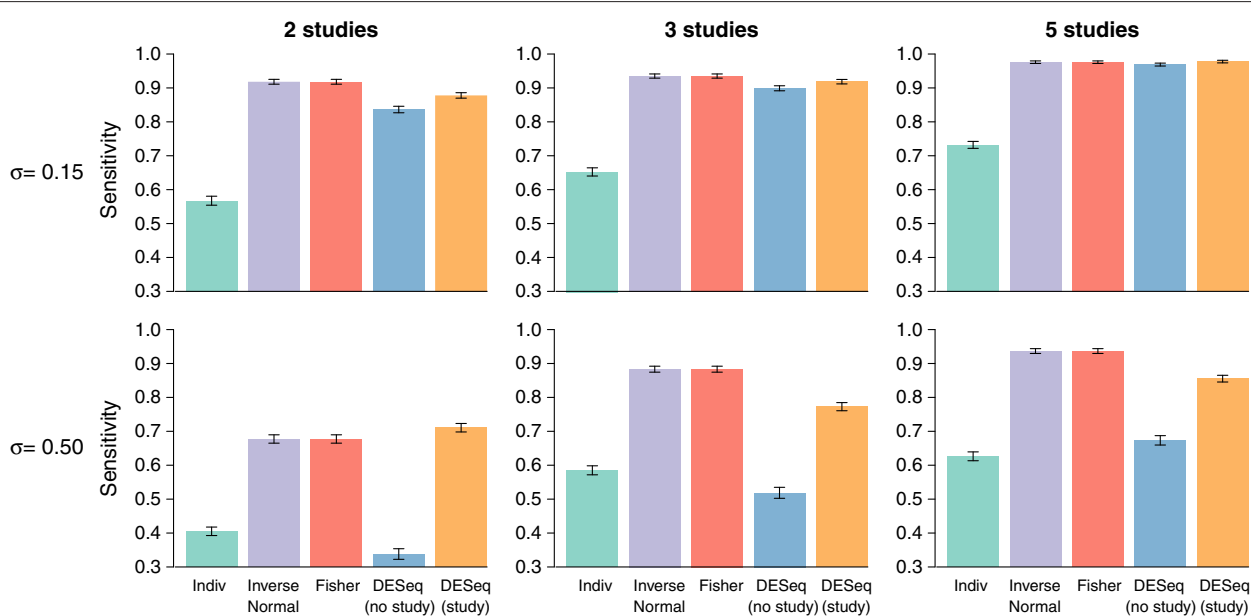
were found to perform very similarly and were able, in the case of large inter-study variability, to outperform the global analysis in terms of AUC (Additional file 1: Figure S1). In particular, in the presence of large inter-study variability, the naive global analysis without a study effect unsurprisingly has the lowest AUC, and the two meta-analysis methods yield a larger AUC than the global analysis with a study effect.

Considering the sensitivity (Figure 4 and Additional file 1: Figure S6), the meta-analyses appear to lead to similar, and in some settings considerably higher, detection power compared to the other methods. We note that in all settings, using the intersection of independent analyses leads to much lower sensitivity, even for low or zero inter-study variability. As for the AUC, the sensitivity was found to be considerably improved for the global analysis when including a study effect in the GLM model, particularly for medium to large inter-study variability. The two meta-analysis methods were found to lead to significant improvements in sensitivity as compared to the global analysis in the presence of moderate to large inter-study variability when three or more studies were considered. However, for the setting that most resembles our real



**Figure 3 Receiver operating characteristic curves for low and high inter-study variability.** Receiver Operating Characteristic (ROC) curves, averaged over 300 datasets. Each plot represents the results of a particular setting, with columns corresponding (from left to right) to simulations including 2 studies, 3 studies, and 5 studies, and rows corresponding (from top to bottom) to simulations with inter-study variability set to  $\sigma = 0.15$  and  $\sigma = 0.50$  (small to large inter-study variability). Within each plot: Fisher (red lines), global analysis with no fixed study effect (DESeq (no study), blue lines), and global analysis with a fixed study effect (DESeq (study), orange lines). The dotted black line represents the diagonal.





**Figure 4 Sensitivity for low and high inter-study variability.** Sensitivity, for the simulation settings corresponding to  $\sigma = 0.15$  and  $\sigma = 0.50$ . Each barplot represents the results of a particular setting, with columns corresponding (from left to right) to simulations including 2 studies, 3 studies, and 5 studies, and rows corresponding (from top to bottom) to simulations with inter-study variability set to  $\sigma = 0.15$  and  $\sigma = 0.50$  (low inter-study variability to large inter-study variability). Within each barplot, from left to right: Individual per-study analyses (green bars), inverse normal (purple bars), Fisher (red bars), global DESeq with no study effect (blue bars), and global DESeq with a fixed study effect (orange bars).

data analysis (2 studies,  $\sigma = 0.50$ ), the global analysis with study effect and meta-analyses appear to have similar detection power. Finally, we also note that for all methods the FDR was well controlled below 5% (Additional file 1: Figure S2).

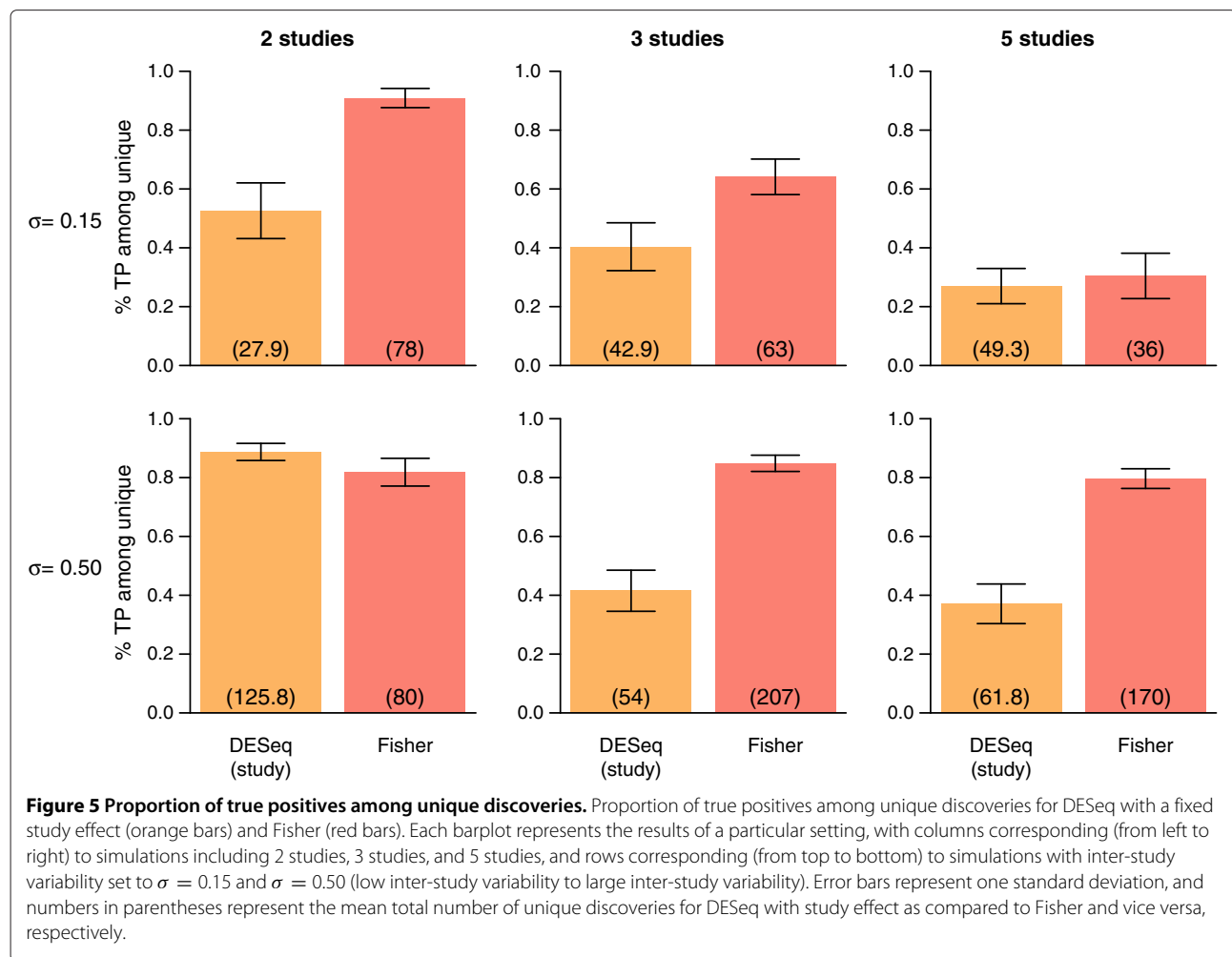
Based on these criteria, the two proposed meta-analysis methods (inverse normal and Fisher) seem to perform very similarly. In order to more thoroughly investigate the differences between  $p$ -value combination methods and the global analysis including a study effect, we calculated the proportion of true positives uniquely detected by the Fisher method as compared to the global analysis with study effect, and vice versa (Figure 5 and Additional file 1: Figure S7). In the setting closest to the real data analysis presented above (two studies and large inter-study variability), the proportion of true positives found uniquely by either the Fisher approach or the global analysis with fixed study effect are very large (around 80% for both methods). This seems to suggest that the additional genes uniquely found either by the global analysis or Fisher  $p$ -value combination method in the real data application may indeed be of great biological interest. For more than two studies, however, as the inter-study variability increases the proportion of truly differentially expressed genes uniquely found by the Fisher method increases compared to the global analysis. For example, for three studies with large inter-study variability ( $\sigma = 0.5$ ), the proportion of truly DE genes uniquely found with the Fisher method was

equal to more than 80%, whereas it was only around 40% for the global analysis with a study effect.

## Conclusions

The aim of this paper was to present and compare different strategies for the differential meta-analysis of RNA-seq data arising from multiple, related studies. As expected, naive analyses such as the overlap of lists of differentially expressed genes found by individual studies or a global analysis not accounting for a study effect perform very poorly. On the other hand, the two proposed meta-analysis methods seem to have very similar performances. For low inter-study variability, the results are very close to those of a global GLM analysis including a study effect. When the inter-study variability increases, however, the gains in performance in terms of AUC, sensitivity, and proportion of true positives among uniquely identified genes for the meta-analysis techniques are significant as compared to the global analysis, particularly for the analysis of data from more than two studies. We note that both of the proposed  $p$ -value combination methods are implemented in an R package called *metaRNASeq*, available on the CRAN; a package vignette describing the use of *metaRNASeq* may be found in Additional file 2 as well as by calling `vignette("metaRNASeq")` after loading the package in R.

Our focus in this work is on differential analyses between two experimental conditions, but can readily be



extended to multi-group comparisons. However, as previously noted, the methods presented here are intended for the analysis of data in which all experimental conditions under consideration are included in every study, thus avoiding problems due to the confounding of condition and study effects. As with all meta-analyses, the  $p$ -value combination techniques presented here must overcome differences in experimental objectives, design, and populations of interest, as well as differences in sequencing technology, library preparation, and laboratory-specific effects.

The differential meta-analyses presented here concern expression studies based on RNA-seq data. However, other genomic data are generated by high-throughput sequencing techniques, including chromatin immunoprecipitation sequencing (CHIP-seq) and DNA methylation sequencing (methyl-seq), and the proposed techniques could potentially be extended to these other kinds of data. However, in order to be biologically relevant, the  $p$ -value combination methods rely on the fact that the same test statistics, or in the case of RNA-seq data conditioned

tests, are used to obtain  $p$ -values for each study. An important challenge for the future will be to propose methods able to jointly analyze related heterogeneous data, such as microarray and RNA-seq data, or other kinds of genomic data. This is not straightforward in a meta-analysis framework and remains an open research question.

## Additional files

**Additional file 1: Supplementary materials.** This document contains a discussion concerning the filtering of RNA-seq data, supplementary information about the characteristics of the data and the Ingenuity Pathways Analysis discussed in the real data analysis, and supplementary figures.

**Additional file 2: metaRNASeq vignette.** This document contains a vignette describing in greater detail the metaRNASeq R package. It can also be obtained by running the following commands in the R console:

```
> library(metaRNASeq)
> vignette("metaRNASeq")
```

## Competing interests

The authors declare that they have no competing interests.

# Authors' contributions

AR participated in the design of the study, performed simulations and data analyses, and helped draft the manuscript. GM designed the study, wrote the associated R package, and helped draft the manuscript. FJ conceived of the study, participated in its design, and drafted the manuscript. All authors read and approved the final manuscript.

# Acknowledgements

We thank Thomas Strub, Irwin Davidson, Céline Keime and the IGBMC sequencing platform for providing the RNA-seq data, as well as two anonymous reviewers for their helpful comments and suggestions.

# Author details

<sup>1</sup>INRA, UMR1313 Génétique animale et biologie intégrative, 78352 Jouy-en-Josas, France. <sup>2</sup>AgroParisTech, UMR1313 Génétique animale et biologie intégrative, 75231 Paris 05, France. <sup>3</sup>Université Lille Nord de France, UDSL, EA2694 Biostatistics, Lille, France. <sup>4</sup>Inria Lille Nord Europe, MODAL, Lille, France.

Received: 13 June 2013 Accepted: 21 March 2014

Published: 29 March 2014

# References

- Marot G, Foulley JL, Mayer CD, Jaffrézic F: **Moderated effect size and P-value combinations for microarray meta-analyses.** *Bioinformatics* 2009, **25**(20):2692–2699. doi:10.1093/bioinformatics/btp444.
- Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and model interstudy variation.** *Bioinformatics* 2003, **19**(Suppl 1):84–90.
- Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple yet powerful new method to detect differential regulated genes in replicated microarray experiments.** *FEBS Lett* 2004, **573**:83–92.
- Hu P, Greenwood CM, Beyene J: **Statistical methods for meta-analysis of microarray data: a comparative study.** *Inf Syst Front* 2006, **8**:9–20.
- Hong F, Breitling R: **A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.** *Bioinformatics* 2008, **24**(3):374–382.
- Tseng GC, Ghosh D, Feingold E: **Comprehensive literature review and statistical considerations for microarray meta-analysis.** *Nucleic Acids Res* 2012, **40**(9):3785–3799.
- Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**(Article 3). http://dx.doi.org/10.2202/1544-6115.1027.
- Jaffrézic F, Marot G, Degrelle S, Hue I, Foulley JL: **A structural mixed model for variances in differential gene expression studies.** *Genet Res* 2007, **89**:19–25.
- Auer P, Doerge R: **A two-stage Poisson model for testing RNA-seq data.** *Stat Appl Genet Mol Biol* 2011, **10**(26):1–26.
- Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(R106). doi:10.1186/gb-2010-11-10-r106.
- Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139–140.
- Kulinskaya E, Morgenthaler S, Staudte RG: *Meta Analysis: a guide to calibrating and combining statistical evidence*, Volume 756 of Wiley Series in Probability and Statistics. West Sussex, England: John Wiley & Sons; 2008.
- Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC Bioinformatics* 2013, **14**:91.
- Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaeffer B, Le Crom S, Jaffrézic F: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Brief Bioinform* 2012. [doi:10.1093/bib/bbs046].
- Stouffer S, Suchman E, DeVinney L, Star S, Williams JRM: *The American soldier. Adjustment during Army life*. Princeton, NJ: Princeton University Press; 1949.
- Liptak T: **On the combination of independent tests.** *Magyar Tudományok. Akademia Matematikai Kutató Intézetének Közleményei* 1958, **3**:171–197.
- Marot G, Mayer CD: **Sequential analysis for microarray data based on sensitivity and meta-analysis.** *Stat Appl Genet Mol Biol* 2009, **8**(Article 3). [doi:10.2202/1544-6115.1368].
- Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc Ser B* 1995, **57**:289–300.
- Fisher RA: *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd; 1932.
- Rau A, Gallopin M, Celeux G, Jaffrézic F: **Data-based filtering for replicated high-throughput transcriptome sequencing experiments.** *Bioinformatics* 2013, **29**(17):2146–52.
- Pearson K: **On a new method of determining "goodness of fit".** *Biometrika* 1934, **26**:425–442.
- Owen AB: **Karl Pearson's meta-analysis revisited.** *Annals of Statistics* 2009, **37**(6B):3867–3892.
- Strub T, Giuliano S, Ye T, Bonet C, Keime C, Kobi D, Gras SL, Cormont M, Ballotti R, Bertolotto C, Davidson I: **Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma.** *Oncogene* 2011, **30**:2319–2332.
- Wickham H: *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer; 2009.
- Chen H, Boutros PC: **VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.** *BMC Bioinformatics* 2011, **12**:35.

doi:10.1186/1471-2105-15-91

Cite this article as: Rau et al.: Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* 2014 **15**:91.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

